

DOCUMENT RESUME

ED 161 898

95

TM 007 479

AUTHOR Malitz, David
TITLE Identification of Predictors of Coding Performance: Selecting Subjects to be Trained to Use Observation Systems.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 78
CONTRACT NIE-C-74-0088
GRANT NIE-G-78-0116
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); For related document, see TM 007 472

EDRS PRICE MF-\$0.83 HC-\$2.06 Plus Postage.
DESCRIPTORS Attitudes; *Behavior Rating Scales; *Classroom Observation Techniques; Codification; *Evaluators; Higher Education; Intermediate Grades; Performance Factors; *Predictor Variables; *Reliability; Screening Tests; Teacher Behavior

IDENTIFIERS Flanders System of Interaction Analysis; *Interrater Reliability; *Minnesota Teacher Attitude Inventory; Observer Reliability; Teacher Evaluation Scale

ABSTRACT

Two predictors of coding performance were validated when 26 trained subjects used the Flanders System of Interaction Analysis to code 11 half-hour videotapes of fifth grade classrooms. The Teacher Evaluation Scale, a high-inference rating system based on the Global Rating Scales, was also used. The predictors were scores on a shortened form of the Minnesota Teacher Attitude Inventory (MTAI-S), and reliability of coding after a few hours of training. Significant correlations with an index of coding performance across the 11 tapes were found for the MTAI-S and for early training reliability of coding. Interest was focused on the MTAI-S as a screening device because of the ease and low cost of its administration. Subjects were divided into two groups, one consisting of high scorers on the MTAI-S (HI-M), the other of low scorers (LO-M). Intraclass correlations were computed for the two groups separately on each of the 11 tapes, and the HI-M group was found to have a significantly higher average intraclass correlation across the 11 tapes than the LO-M group. These findings were discussed in terms of their implications for screening and predicting; and for understanding the nature of observational coding systems.
(Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED161898

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SCOPE OF INTEREST NOTICE

The ERIC Facility has assigned this document for processing to:

TM IR

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Identification of Predictors of Coding Performance:
Selecting Subjects to be Trained to Use Observation Systems

DAVID MALITZ

Research and Development Center for Teacher Education
The University of Texas at Austin

TM007 479 624

This study was supported in part by the National Institute of Education Contract No. NIE-C-74-0088, Grant No. OB-NIE-G-78-0116, Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred. Paper presented at the annual meeting of the American Educational Research Association held in Toronto, March 27-31, 1978.

Abstract

In order to identify predictors of coding performance which could be utilized to screen coders, 26 trained subjects used the Flanders System of Interaction Analysis to code 11 half-hour videotapes of fifth grade classrooms. Predictors which were analyzed included scores on a shortened form of the Minnesota Teacher Attitude Inventory (MTAI-S) and reliability after a few hours of training. Significant correlations with an index of coding performance across the 11 tapes were found for the MTAI-S ($r = .45$) and for training reliability ($r = .48$). Interest was focused on the MTAI-S as a screening device because of the ease and low cost of its administration. Subjects were divided into two groups, one consisting of high scorers on the MTAI-S (HI-M) ($n = 14$), the other of low scorers (LO-M) ($n = 10$). Intraclass correlations were computed for the two groups separately on each of the 11 tapes, and the HI-M group was found to have a significantly higher average intraclass correlation across the 11 tapes ($r_I = .88$) than the LO-M group ($r_I = .79$) ($p < .05$). These findings were discussed in terms of their implications for screening and prediction and in an effort to understand the nature of observational coding systems.

Identification of Predictors of Coding Performance:

Selecting Subjects to be Trained to Use Observational Systems

Researchers who intend to conduct studies in which systematic observational data are to be collected are often faced with the practical managerial problems of training coders. Those who train people to use observational coding systems realize that, as in any learning situation, some people will learn to use the system more accurately than others. Obviously, the better the coder, the greater the reliability of his or her coding and the more meaningful the data, results, and conclusions derived from that coding. Because of the importance of obtaining good data from coders in any study depending upon systematic observation, it would be useful to be able to identify either before or toward the beginning of training, candidates who are most likely to be accurate coders. If predictors of coding performance could be isolated, people most likely to code well could be selected out of a larger group. In this way training time and effort would be spent on those most likely to perform well and the overall quality of the data obtained from the study's coders would be improved over what would be expected from an unscreened group of coders.

Potential predictors of coder performance emerged from discussions among staff members of the Evaluation of Teaching Program (EOTP) of the Research and Development Center for Teacher Education at The University of Texas at Austin. These staff members had, in the course of previous studies, trained approximately 35 coders to use the Flanders System of Interaction Analysis (FSIA) (Flanders, 1970). In addition, several staff members had themselves been trained to use the FSIA as well as other

classroom observation systems. On the basis of this experience, staff members identified two variables which they felt might be potentially useful in screening future coders.

The first of these predictors was coding performance after several hours of training. Typically, when training a new group of coders, trainers will begin with a general introduction to the use and nature of the observation system. Following this, coders are introduced to each of the categories on the system, and the behaviors covered by each of the categories are defined. Once the categories have been explained, trainers usually have subjects code a sample of behavior (either from audio or videotape). The trainees' codings of this sample are discussed and feedback is given on their performance. For the purposes of prediction, this first reliability check, as it is termed, can be viewed as a sample of on-the-job performance. Of course, performance on the first reliability check is not totally representative of later performance since at this point in training subjects have had limited exposure to the system and are using it imperfectly. However, even at this early point in training, individual differences do exist among coders in their coding performance, and it was hypothesized that this variability would reflect differences in skills necessary to learn and use an observation system accurately.

The second predictor variable identified by the EOTP staff was coders' attitudes towards teachers, students, and classrooms. Because no observation system can capture completely all of the complexities of classroom behavior, ambiguities frequently exist within these systems. In order to compensate for these ambiguities, a good deal of the training process involves alerting coders to typical classroom behaviors which do not easily fit into the classificatory scheme of the system in use. Often,

somewhat arbitrary ground rules are established in training which are designed to aid coders in deciding how to code a particular behavior which imperfectly matches the categories available. However, it is impossible for ground rules to be established which will pertain to every situation which coders will encounter while observing classroom behavior, and for this reason coders frequently must rely upon their own judgment in coding behaviors which cannot be coded unambiguously. It is in these ambiguous situations that coders' attitudes towards teachers and students might come into play.

An example of such an interplay between attitudes and coding can be seen in coding the behavior "teacher criticism," a behavior found on many classroom observation systems. When a coder observes a teacher correcting a student, he must decide whether this behavior contains the affective components necessary to classify it as criticism rather than as some alternative behavior (such as "teacher correction--academic" or "gives directions"). If the situation is an ambiguous one and if no ground rules have been learned which can clarify it, the coder must rely upon his own experience and background to decide whether the teacher's behavior qualifies as criticism. One way in which a coder might make such a decision is in terms of his attitudes about what happens and ought to happen in classrooms. If the coder believes that teachers ought to be gentle with students, he might be apt to classify a particular teacher behavior as "criticism" whereas another coder who believes that teachers should be strict might not see the teacher as critical.

If ambiguous situations occur frequently enough and if coders' attitudes help them resolve these ambiguities, then one would expect

that a relationship would exist between measures of these attitudes and coding performance. The exact nature of this relationship was not clear to the EOTP staff, but it was felt to be worthy of investigation.

Thus, on the basis of experience with teacher observation systems, two variables were identified as potential predictors of coding performance. These two variables were coding performance early in training (a behavioral measure), and attitudes towards teachers and pupils (a paper and pencil measure). The opportunity to investigate these variables came when a study was planned to explore the nature of "coder drift": the tendency of coders who code together to resolve ambiguities in a similar manner and for coders who code alone to resolve them in idiosyncratic ways. In this study (Marston, Zimmerer, & Vaughan, 1978), a group of coders was trained to use the Flanders System of Interaction Analysis (FSIA). Once trained, they viewed 11 videotapes of classroom behavior either with or without a partner in the coding room. A within-subject design was employed in which each coder saw five of the videotapes with his partner in the coding room and five other videotapes in isolation. An 11th tape was viewed by each subject in the middle of the experiment. This tape was the same tape viewed at the end of training and was included in order to measure test-retest reliability. The order of the experimental tapes as well as the order of the together-apart manipulation was randomized across pairs.

This design allowed investigation of predictors of coding performance on the experimental tapes. An instrument measuring attitudes toward the classroom was administered when coders were hired and coding performance was measured at two points in training. This provided the data necessary to assess the usefulness of coder attitudes and training performance as predictors of coding performance.

Method

Coders

Twenty-eight paid student volunteers were recruited mainly through advertisements at the financial aids office. They were paid \$4 per hour of experimental participation with a \$20 bonus for completion of the test series. Because two coders had insufficient time periods available to be scheduled to see the tapes, only 26 coders completed the experiment.

The 26 coders who completed the study ranged in age from 18 to 34, with an average age of 24.5. The sample included 12 males and 14 females. Approximately half of the subjects were undergraduates and the other half were graduate students. Six of the coders majored in education, five in psychology, and the rest in various arts and sciences. Responses on a questionnaire administered before training indicated that half of the subjects had worked at some time as a teacher, although only a third had had practice-teaching experience. No subject indicated that he or she knew how to use any classroom observation systems.

Observation Systems

Two observation systems were used in this experiment. The first was the FSIA, a widely-used classroom observation system. This system focuses upon verbal interaction in the classroom and contains ten categories. Coders are required to observe classroom interaction and to mark one of the ten categories every three seconds. The data resulting from an observation period are frequencies of tallies in each of the ten categories. In order to correct for differences in coding speed among coders, each coder's category frequencies are divided into the total number of codes he made for the time period. This manipulation converts

category frequencies to proportions which are directly comparable across coders.

The second observation instrument used was the Teacher Evaluation Scale (TES), a high-inference rating system based on the Global Rating Scales (GRS) (Poynor & Wetherold, 1976). The GRS was developed by the Evaluation of Teaching Program staff after studying Rosenshine's (1971) review of research studies relating teacher behavior to pupil achievement and then reviewing existing teacher observation instruments. For the purposes of the present study, two members of the EOTP staff modified the GRS by the process of reviewing the categories and then dropping those that were felt to be redundant, retaining those that seemed applicable, and adding categories that would make the instrument more comprehensive. Wording of the definitions of categories was also changed in those cases where it was felt a clearer description of the variable to be rated was needed. A copy of the TES is included as Appendix A.

The TES was used as a high-inference companion instrument to the FSIA. After completing the observation period for a given videotape, coders rated the teacher on each of the nine TES categories. These ratings indicated the amount or level of behavior observed on a five-point Likert-type scale. These ratings were used directly for all analyses involving the TES.

Training

Training in the use of the FSIA was given in two 3 1/2 hour sessions on consecutive evenings. The training was based on the audio cassette

method provided by an expert FSIA trainer.¹ Some minor modifications in this method had to be made to accommodate the large group being trained. In particular, it was not possible to give any feedback on the coders' reliability since they did not have partners at this time. Because of this, considerable emphasis was given to going over the correct coding for the training tapes. At the end of each training session the coders made an actual coding of a tape from the same series that they would be using for the test part of the experiment. Reliabilities between all possible pairs of coders were computed for these two tapes and averaged. These reliabilities were then used to assign coders to pairs so that all pairs were approximately equal in the training.

There was no formal training in the use of the TES. Coders were given copies of the instrument and were told to examine the category descriptions. However, no attempt was made to explain the categories at great length or to solicit questions about the category definitions.

Materials

The coders viewed each experimental videotape in one of two office-sized rooms equipped with a 48 cm (19 inch) black and white television monitor, a microphone, a button to signal the experimenter, and a work table. Under the isolation or "separate" condition, one coder worked in each room while under the "together" condition either a single pair or two pairs of coders worked at the same time.

The experimenter presented the videotapes using a Sony VP-2000 U-matic 19mm (3/4 inch) cassette videoplayer. Tone signals were recorded directly onto the videocassettes which marked the end of 10-minute time intervals and which indicated that coders were to start a new FSIA coding sheet. An audiotape recording was made of everything the coder said from the time the coder entered the coding room until he or she left. Although coders were not told that they were being recorded, most realized by the end of the experiment that they were.

The tapes were selected from a series done for the Individualized Teaching for Effective Coping Study². These tapes were 30-minute segments of either math or social studies lessons in an actual fifth grade class using traditional instructional methods. Two math lessons were selected for the reliability training tapes and the best quality social studies lesson for 10 other teachers were used for the test tapes.

Method and Design

The coders were assigned to code the experimental tapes in one of two groups. In the first group they worked with their partner for the first five tapes and then worked in separate rooms for the last five tapes. In the second group, their partner came in at the same time but they worked in separate rooms for the first five tapes and then worked together for the last five tapes. The coders were encouraged by the experimenters to be as accurate as possible but were never told that they should maintain a high reliability with their partner. If the coders asked any questions about the FSIA system they were told to refer to the materials they had received during the training. Any questions about

the experiment were deferred until the debriefing session after the last tape was coded.

The coder pairs were formed using the reliabilities from the two training videotapes and schedules of available times provided by the coders. This was done by first calculating a Scott's coefficient (Scott, 1955) between every possible pair of coders for each of the two training tapes. These reliabilities were then transformed using the arcsin method for proportions and averaged. Finally, a trial-and-error procedure was used to find a set of pairs with compatible schedules that had training reliabilities close to the overall average and that showed the least variance in reliability between pairs. Seven pairs were randomly assigned to each of the groups and scheduled so that with the exception of a few cases they would never see more than one tape a day. One pair had to be dropped after the first six tapes in the together-apart group.

Because earlier work with these tapes indicated that there were large differences in reliability among them, seven random orders for the 10 tapes were generated and one pair in each group was assigned to the first six orders. The seventh order was given to the extra pair in the apart-together group.

In addition to the sessions required to code these 10 tapes, an extra session was scheduled after the first five sessions. During this extra session subjects coded for a second time the tape they had seen on the second day of training so that a test-retest reliability could be computed for each coder. The coding for this tape was done under the same condition as the first five sessions. That is, coders who had

been working together saw the reliability tape together and vice versa. The coders then switched to the opposite condition for the remaining five sessions. The test sessions took 17 days to complete.

Attitude Instrument

Numerous measures of attitudes towards teachers, students, and classrooms were examined in order to find one suitable for this study. A number of instruments were eliminated because they were outdated, ambiguous, or poorly constructed. The Minnesota Teacher Attitude Inventory (MTAI) (Cook, Leeds, & Callis, 1951) was finally chosen because it seemed to most adequately sample a variety of attitudes which might be relevant to the classroom behaviors to be observed by the coders. Unfortunately, the MTAI, containing 150 items, was considered too long to be completed by coders in the time allotted before training began. Therefore, an instrument of between 30 and 40 items was constructed which consisted of a subset of MTAI items and which could be completed in a short period of time. The 150 items in the original instrument were closely examined and those which appeared ambiguous, irrelevant, or obsolete were eliminated. The remaining items were grouped by content areas (e.g., discipline, student initiative, etc.) and items representative of each of these groups were chosen to be on the final instrument. This final instrument, a shortened version of the MTAI (MTAI-S), consisted of 35 items which were felt to adequately sample classroom attitudes for the purposes of this study. It was this shortened version which was administered to all coders before training began. The items contained in the MTAI-S are listed in Appendix B along with their original item numbers on the MTAI.

Data were obtained from a separate study which allowed examination of the relationship between the MTAI-S and the MTAI. In that study, the MTAI had been administered to 365 candidates for a student teaching program in Canada. Protocols were obtained for these subjects, and 295 were identified which had no missing data. Scores on the full MTAI as well as on the 35 MTAI-S items were obtained from these protocols, and a Pearson product-moment correlation was computed between these two sets of scores. A correlation of .89 ($p < .001$) was obtained between the MTAI and MTAI-S scores, providing strong evidence that the 35 items chosen for use on the MTAI-S were, in fact, highly representative of the MTAI as a whole. A description of the MTAI data base and of the procedures used to analyze it are provided in Appendix C.

Results

Twenty-eight subjects completed training in the FSIA. However, one of these coders moved out of town before the experimental portion of the study was begun. Therefore, this coder as well as the person who had been paired with him had to be dropped from the study, leaving 26 coders. All of these coders completed the study, and all analyses are based on data from these 26 subjects.

In order to predict coding performance on the 11 experimental tapes, an index of this performance had to be developed. Ideally, it would be best to measure each coder's agreement with standard codes for each of the videotapes. This would be an acceptable procedure if such standard codes, known to be "correct," could be obtained. However, because of concern with the ambiguities in the coding of certain behaviors, the EOTP staff did not consider this an acceptable strategy. Although

expert coders might be able to agree on how to code much or perhaps most of the behavior on a particular tape, there undoubtedly would be disagreements about certain behaviors. Since it was felt that these ambiguous behaviors might be the very ones which would produce variance which could be accounted for by coders' attitudes, coder performance was not measured against an "objective" coding standard.

Instead, an index was developed which reflected coders' average agreement with all other coders. This index, in the case of the FSIA, was based upon Scott's coefficient (Scott, 1955), a frequently-used index of interrater agreement. For a given videotape, each subject's codes for the 30 minutes of behavior were tallied, yielding frequencies for each of the 10 categories of the FSIA. After category frequencies were obtained for each of the 26 coders on a given videotape, Scott's coefficient was computed for each pairing of the 26 coders and these coefficients were placed in a matrix. Once organized in matrix form, an average was computed for each coder which represented that coder's mean Scott's coefficient with the other 25 coders in the sample. This index, termed "average reliability" in this paper, reflects the degree to which each subject's codes agreed with the codes of the other 25 coders. Average reliabilities were computed in a similar manner for the TES data, as will be discussed later in this paper.

As a measure of coder performance, average reliability suffers from the fact that it is not measured against an objective standard of "good" or "correct" coding. Instead it is a bootstrap measure in which the best coder is defined as the one who achieves the highest reliability with all other coders. However, from a trainer's point of

view, this is a very useful index of performance, for a person who scores high on this index is one who, if randomly paired with another coder from the same training group, would likely achieve high reliability. Thus, this index reflects the degree to which a coder is a "good" coder, from the standpoint of interrater agreement.

For both the FSIA and the TES, average reliability coefficients were computed for each coder on each of the 11 videotapes. These coefficients were themselves averaged across all of the tapes, reflecting each coder's average performance during the experimental portion of the study. These coefficients served as the criterion measures.

Average reliability coefficients were also computed for the first and second reliability checks taken during the training sessions. In addition, an average was computed across these two reliability checks which reflected each coder's average performance during training. These three coefficients were used to predict criterion performance from training performance.

The 35 items of the MTAI-S were scored in the manner prescribed by the scoring keys for the MTAI (Cook, et al., 1951). This score, a measure of coder attitudes, was also used as a predictor variable.

Analyses involving the FSIA will be considered first; the TES data will be dealt with later in this paper. As an initial step in assessing the relationships between the predictors and the criteria, Pearson product-moment correlations were computed among the MTAI-S scores, the average reliabilities for the first and second reliability checks, the average reliability for training as a whole, and the average reliability on the FSIA across the 11 experimental tapes. These

correlations are presented in matrix form in Table 1.

Insert Table 1 About Here

Examination of the first row of this table reveals the correlations of the predictors with the criterion. It can be seen that all predictor variables correlated significantly ($p < .05$) with the criterion. The reliability after a few hours of training correlated more highly with the criterion than reliability later in training and almost as well as the reliability summed across the two reliability checks. From this data it would appear that a coder's reliability early in training is as good a predictor of his final performance as is his reliability later in training. Somewhat surprising, however, is the fact that subjects' scores on the MTAI-S, a 35 item paper-and-pencil test given before any training, predicts performance as adequately as the behavioral measures.

Further examination of Table 1 shows that the MTAI-S is significantly correlated with the training reliabilities, although these correlations are not quite of the magnitude of its correlation with the criterion. In addition, it can be seen that training check one is significantly correlated with training check two although not as highly as with the criterion. Finally, it should be noted that subjects' experimental group has been included as a variable in Table 1. For this variable, subjects who coded with their partner during the first half of the experiment and separately from their partner on the second half received a code of "1." Subjects in the apart-together group received a code of "2." Analysis of the experimental data for this study (Marston, Zimmerer, & Vaughan, 1978) revealed that regardless of experimental group, subjects attained higher reliability when coding together than when coding apart. However, a

main effect was found for experimental group as well, such that the together-apart group coded more reliably across conditions than the apart-together group. This fact is reflected in the significant negative correlation in Table 1 between experimental group and the criterion measure of coding performance across the 11 experimental tapes. It should be noted, however, that none of the predictor measures was significantly correlated with the experimental group variable.

The above data is encouraging for the search for predictors of coding performance. It appears that the paper-and-pencil measures of coding performance predicts almost as well as the behavioral measure of reliability during training. In addition, it appears that none of these predictors was confounded with the experimental manipulation which comprised the main portion of this study. In order to further investigate these relationships, linear regression models were constructed to test the effects of each of these variables in predicting the criterion. In all of these models, the average reliability on the FSIA across the 11 experimental tapes served as a criterion vector.

As a first step in this regression analysis it was necessary to ascertain the effects of the experimental manipulation which comprised the major portion of this study, upon the relationships between the predictors and criterion. Although the present investigation was superimposed upon the main study, it was hoped that the relationships between attitudes and coding performance could be studied independently of the manipulation. In terms of statistical relationships, the question to be answered was: when the effects of the predictor variables were

controlled for, was there a difference in the level of performance for the two experimental groups? This question was answered by use of regression models which may be recognized as analysis of covariance models. In these models, the attitude and training performance variables served as covariables, and tests were made for an interaction between the experimental manipulation and the criterion. In order to simplify these models, only data from the first reliability check were used as a measure of training performance.

The full model which was constructed to answer this question was as follows:

$$(1) \quad Y = a_0 + a_1X_1 + a_2X_2 + a_3(X_1 \cdot A) + a_4(X_2 \cdot A) + a_5(X_1 \cdot T) + a_6(X_2 \cdot T) + E_1$$

where

Y = a vector containing the average reliabilities on the FSIA across the experimental tapes,

X_1 = a vector containing "1" for subjects in the apart-together experimental group, "0" otherwise,

X_2 = a vector containing "1" for subjects in the together-apart experimental groups, "0" otherwise,

A = a vector containing MTAI-S (attitude) scores,

T = a vector containing average reliabilities for the first training check.

The coefficients a_0 through a_6 are regression coefficients resulting in a least squares solution for the equation, and E_1 is the residual vector. Note that the effect of the multiplicative vectors is to produce vectors which contain either the attitude or training score for subjects in a particular group, and zeroes for subjects in the other group. These multiplicative vectors allow the regression lines between the predictors and the criterion to take on different slopes for the different experimental groups. In other words, such a model allows for an experimental group interaction across levels of attitudes and training performance.

Restrictions may be placed upon model 1 such that $a_3 = a_4$ and $a_5 = a_6$. Such restrictions result in model 2:

$$(2) \quad Y = b_0 + b_1X_1 + b_2X_2 + b_3A + b_4T + E_2$$

The notation for the variables in model 2 is the same as for those in model 1, except that the regression coefficients are now represented by b's rather than a's to emphasize the fact that the least squares solution for this model will be different from model 1. This restricted model does not allow for different group slopes for the predictors of interest, but assumes instead that group membership need not be considered in describing the slope of these regression lines. Thus, if model 2 can be treated as a true model, then the relationship between the attitude and training performance variables and the criterion can be treated and discussed without regard to the experimental manipulation which comprised the main portion of the study.

Models were constructed in which the two restrictions discussed above ($a_3 = a_4$ and $a_5 = a_6$) were imposed one at a time and in which they were imposed

simultaneously, resulting in model 2. In none of these three occasions was a significant F ratio ($p < .05$) obtained. Thus, these tests argued in favor of the use of model 2 as a true model and allowed further analysis to be conducted without regard to the experimental manipulation.

Using model 2 as a full model, it was discovered that a further simplifying restriction could be placed upon the model such that $b_1 = 0$ and $b_2 = 0$. This eliminated the need for considering different intercepts for the two experimental groups. Thus, the resulting simplified model was:

$$(3) \quad Y = c_0 + c_1A + c_2T + E_3$$

When model 3 was used as a restricted model and model 2 as the full model, the resulting F ratio was non-significant ($p > .05$), allowing model 3 to be used as the true model. Note that this model is a considerably simplified version of the beginning model (model 1) and that this simplification did not result in a significant increase in prediction error.

Using model 3 as the full model, further questions could be investigated. From a practical point of view, an interesting question is whether both the attitude and the training predictor are necessary to adequately predict coding performance, or whether one or the other will suffice without a significant increase in error.

If only the attitude measure is used (i.e., $c_2 = 0$), then model 4 results:

$$(4) \quad Y = d_0 + d_1A + E_4$$

If, on the other hand, only the training predictor is used, then model 5 results:

$$(5) \quad Y = e_0 + e_1T + E_5$$

F tests were performed using model 3 as the full model and models 4 and 5 alternately as the restricted model. When model 5 was tested, the resulting F ratio was not significant ($p > .05$), but when model 4 was tested, the probability was marginally significant ($p = .0546$, with 1 and 23 degrees of freedom). Thus, it can be stated with some confidence that while the attitude measure can be eliminated from the prediction equation with no significant increase in error, the same cannot be said of the training reliability measure. If one relies solely on the attitude measure, it is at the expense of a marginally significant loss in predictive efficiency.

To summarize the results of the analyses described above, it was discovered that both the attitude measure and the training predictor were significantly correlated with coding performance using the FSIA on the experimental tapes. These relationships were found to be independent of the experimental manipulation which constituted the main portion of the study. Although each of these two predictors was significantly related to the criterion, it was found that training performance accounted for more of the variance, since relying upon the MTAI-S alone resulted in significantly more prediction error than using both predictors. Using both predictors (model 3), a multiple R of .5636 was obtained, as compared with r's of .4802 for training performance alone and .4460 for the MTAI-S alone.

Although training performance accounts for more of the variance than the attitude score, a trainer in a practical situation might be more interested in using the attitude score since it is more easily obtained. As a screening device the MTAI-S has the advantage of requiring minimal expenditure of time or resources to be administered. Thus, it is of interest to see how effectively this instrument might be used as a screening device.

In order to do this, subjects were divided into two groups, one consisting of subjects scoring high on the MTAI-S (HI-M), the other of low scorers (LO-M). The cutoff score for this division was determined by examining the distribution of MTAI-S scores for the 26 coders. Scores on the MTAI-S were found to range from -4.0 to 18.0 with a mean of 6.2. A break in the distribution where no scores were found was located near the mean of the distribution, so this was used as the cutoff point. The 12 subjects with MTAI-S scores less than 6.0 were placed in the LO-M group, while the 14 subjects with scores greater than or equal to 6.0 were placed in the HI-M group. In order to compare the coding performance of these two groups, intraclass correlations were computed across the 10 FSIA categories for each of the 11 experimental tapes. An unanchored intraclass coefficient (Winer, 1971) across all raters in each group was used. The results of this analysis are depicted in Figure 1 in which the intraclass coefficients for each group are shown for each of the tapes.

Insert Figure 1 About Here

It can be seen that in all cases (except for tape 11) the HI-M group attained higher reliability than the LO-M group. A t-test was performed on these reliabilities, and the two groups were found to be significantly different ($t = 3.53$, $df = 10$, $p = .005$). Across the 11 tapes, the HI-M group achieved an average intraclass coefficient of .8820, while the LO-M group achieved a considerably lower coefficient of .7931. Furthermore, examination of Figure 1 shows the HI-M group to have achieved more consistent reliabilities across the tapes while the LO-M reliabilities fluctuate considerably from tape to tape. From this data it is apparent that the MTAI-S can be used to divide coders into groups which will differ

considerably in the reliability they will achieve after training. Of course this instrument requires cross-validation on other samples and in different situations before it can be considered a valid screening device. The data available from the coding of the TES allowed such a cross-validation. Although the TES data is not independent of the FSIA data (since the same coders coded both instruments), a replication of the above findings using the TES would lend greater credence to the relationship between coder attitudes and coding performance.

The TES data were analyzed in a manner analogous to the FSIA data. Average reliabilities were first computed, which were then correlated with the MTAI-S scores. However, there was some question as to what measure of reliability was appropriate as the basis for the average reliability matrices. Scott's coefficient, which was used for the FSIA, was considered inappropriate for the Likert-type ratings on the TES. Four other indices of reliability were used, and average reliabilities were computed using these various indices, so that their results could be compared. Three of these indices are standard statistics: Pearson-product moment correlations, unanchored intraclass correlations, and anchored intraclass correlations. The fourth index was the percent of categories in which agreement was observed between two coders (agreement was defined as ratings which differed by no more than one point on the five-point rating scale). This last index was termed, simply, "percent agreement."

Average reliabilities were computed across the eleven tapes for each coder using each of these indices. The intercorrelations among the resulting sets of scores were computed and are presented in Table 2.

Insert Table 2 about here

Examination of Table 2 reveals that all measures are highly (and significantly) correlated with one another. It should be noted, however, that the three correlational measures are more highly related to one another than to the percent agreement measure. Furthermore, examination of the means for the four measures (also presented in Table 2) shows that reliability was rather poor on the TES, regardless of the measure used. Even the highest correlational measure (the Pearson product-moment correlation) indicates that on the average, only about 25% of the variance in one coder's ratings could be accounted for by the ratings of the other coders in the sample. Viewed another way, the percent agreement index indicates that on the average, coders agreed on only 75% of the TES categories, disagreeing by more than one out of the five scale points on the remaining 25% of the categories.

Despite these low reliabilities, correlations were computed between the four indices of average reliability and the MTAI-S scores. These correlations are presented in Table 2. It can be seen that the percent agreement index correlated significantly with the MTAI-S, but that the other three measures did not.

This significant relationship between the percent agreement index and the MTAI-S was examined for an interaction with the experimental manipulation using regression models analogous to those used for the FSIA data. No significant increase in prediction error was observed when the interaction term was dropped from the full model and it was therefore concluded that the relationship between the percent agreement index and the MTAI-S scores could be dealt with outside of the context of the main experiment.

It will be recalled that in the analysis of the FSIA data, subjects were placed into high- and low-MTAI groups (HI-M and LO-M) on the basis of their MTAI-S scores. It was shown, using the FSIA data, that if intraclass coefficients were computed separately for the two groups, that the HI-M group achieved higher reliability across the eleven tapes than the LO-M group. A similar analysis was performed using the TES data, but this time no significant difference in the intraclass coefficients for the two groups was found across the tapes. The HI-M group had an average intraclass coefficient across the eleven tapes of .4692, compared with .4887 for the LO-M group. Thus, it can be concluded that screening subjects on the basis of their MTAI-S scores would have made no difference in the TES reliabilities achieved on the experimental tapes.

Discussion

Data have been presented which demonstrate that coder performance is in part predictable from scores on the MTAI-S and from performance after a few hours of training. These relationships are, of course, important from the standpoint of prediction. But in addition to this, the findings of this study may have important implications for understanding the processes of coder training and performance. Issues relevant to both of these points of view will be discussed.

Turning first to prediction, one can say simply that both the MTAI-S and training performance show promise as predictors of coder performance. That training performance is a predictor of later performance is perhaps not too surprising, in view of certain well-established findings in the areas of learning and industrial psychology.

The training process can be viewed in most basic terms as a learning situation. Coders must learn the theoretical orientation of the coding system, the definitions of the categories and the mechanics of encoding observed behavior. It is widely assured among learning psychologists that measures of early learning on a task are predictive of performance on later trials of that task. Thus, with coder training viewed as a learning situation, it would be expected that a sample of performance taken early in the learning process would be predictive of performance after training was completed.

Viewed another way, this sample of early performance is functionally identical to the "job sample" used in industrial psychology. In many situations, the best predictor of how a worker or subject will perform at a particular task tomorrow is how he or she performs that task today (Strauss & Sayles, 1967). Since the job of coding videotapes during training requires behaviors and skills identical to coding tapes during the experiment, one could justifiably expect a similar relationship between this "job sample" and later performance.

Thus, whether viewed as a measure of early learning or as a job sample, one could justifiably expect that performance during training would predict later coder performance. One might not feel as confident, however, in expecting scores on a paper-and-pencil attitude measure to predict such performance. One reason for this is that researchers tend to assume that observational coding systems are "objective," especially a "low-inference" system like the FSIA, which is rather behaviorally oriented. While recognizing that coders might have attitudes regarding the subject matter they are to observe, one hopes that they can "leave

them behind" and observe "objectively." Thus, in an ideal situation there should be no relationship between attitudes and coding performance.

Nevertheless, such a relationship was found on the FSIA. As was shown, these relationships were strong enough so that, in this study at least, the MTAI-S could have been used as a screening device to obtain a group of coders who would have attained higher average reliability than would an unscreened group. The strength of the relationship between the MTAI-S and coder performance was of almost the same magnitude as training performance, but MTAI-S scores are much more readily and inexpensively obtained than measures of training performance. For this reason, the MTAI-S holds promise as a practical and useful screening device which can be applied to large groups of coding candidates with a minimal investment of time and money.

The above comments apply only to the findings involving the FSIA, for a relationship between the MTAI-S and coding performance was not supported by the TES data. A significant correlation between these two variables was found in only one of the four indices of TES coder performance explored, and when coders were grouped as high- and low-MTAI scorers, no significant difference in coding performance was found between the two groups.

Perhaps the reason for this negative finding is the poor reliability among all coders on the TES. It will be recalled that subjects received no formal training in the use of the TES: they were merely handed a description of the instrument and its categories and told to read these materials. Judging from the reliabilities obtained on this instrument, there was little agreement among coders on how to rate teachers on the

categories. Lack of clarity as to category definitions may have introduced a sizable amount of randomness to coders' ratings, and this may have obscured any relationships which might have existed between the MTAI-S and the TES.

Turning from the issue of prediction alone, the findings of this study may have wider implications for understanding the nature of observational coding systems in general. More specifically, while it is an interesting and perhaps useful empirical fact that the MTAI-S scores were related to FSIA coding performance, it is also of interest to speculate about the reasons for this relationship.

In the introduction to this paper, a model of coder behavior was outlined which could account for a relationship between coder attitudes and performance. It was proposed that no classroom observation system could unambiguously classify all behavior which might be observed. It was hypothesized that when a coder encounters behavior which does not clearly fit into any one category, attitudes and expectations about classroom behavior might mediate the forced-choice decision with which the coder is faced, thereby influencing the codes. This model would account for the relationship between the MTAI-S, purportedly a measure of classroom attitudes, and coder reliability. Presumably, coders with similar attitudes would tend to be influenced in their coding decisions in a similar manner and would therefore produce similar coding profiles and achieve high reliability with one another.

However, close scrutiny of this model indicates that it may not adequately explain the data in this study. The assumption of the model is that coders with similar attitudes tend to code in a similar manner, and that this coding similarity will be reflected in high reliability.

It was, in fact, found that coders who scored in the upper half of the MTAI-S distribution did have high reliability (agreement) among themselves.

A reading of the MTAI Manual (Cook et al., 1951) reveals, however, that the MTAI was designed as a bipolar scale. Because of the high correlation (.89) between the MTAI-S and MTAI, considerations which apply to the MTAI should apply to the MTAI-S also. From Cook et al.'s discussion, high scorers on the MTAI are those who believe teachers should be democratic and permissive, while low scorers believe in autocratic, well-controlled classrooms. Thus, just as high scorers on the MTAI (and, by inference, the MTAI-S) can be characterized by a definite set of opinions, so can the low scorers. Therefore, if coders with similar attitudes tend to code in a similar manner, then one would expect coders scoring low on the MTAI-S to code in a homogeneous manner, just as the high scorers did. Although one would not expect the codes of the two groups to coincide, one would expect homogeneity within the two groups. This homogeneity should have been expressed as high within-group reliability. However, it will be recalled that when coders were split into high- and low-MTAI groups, the high-MTAI group was found to have significantly higher intraclass reliability across the 11 tapes.

These findings, therefore, seem to be inconsistent with the model of coding proposed above. An alternative model can be designed which does account for the FSIA findings. This model, which will be termed the "training model" to contrast it with the "attitude model" discussed above, proposes that the MTAI-S, rather than measuring classroom attitudes per se, actually is a measure of conformity or social malleability. According to this model, high scorers on the MTAI-S are seen as conformers,

while low scorers are seen as more nonconforming. The training model hypothesizes that the high scorers, the conformers, were more willing to accept the assumption, attitudes, biases, and somewhat arbitrary ground rules inherent in the training process. Thus, the high scorers basically accepted the orientation towards observing behavior which was being taught during training. With this common grounding and set of attitudes, the high MTAI group proceeded to code the experimental tapes in a relatively homogeneous manner.

Subjects who scores low on the MTAI-S, on the other hand, are seen as less accepting of the attitudes and viewpoints conveyed during training. They tend to rely more upon their own beliefs, to do things their own way, and therefore to code more individualistically, achieving relatively low reliability with other subjects coding in their own idiosyncratic fashion.

This training model was proposed after a close examination of the items on the MTAI-S. These items appear to be very easily faked. It would seem to be an easy task to respond to the items in such a way as to produce a score indicative of permissive, democratic attitudes, on the one hand, or of controlling, autocratic attitudes on the other. Subjects who endorse the democratic, permissive attitudes (considered more socially desirable within a liberal, university atmosphere) can be seen as those most conforming and most open to the liberal educational attitudes implicit in the FSIA system and training program.

As incidental piece of information supports the training model. The expert FSIA trainer who initially trained the EOTP staff (see Footnote 1) indicated that in his experience in training several hundred coders, the worst coding candidates were experienced teachers. This is because

experienced teachers have developed such strong attitudes about education that they have difficulty accepting a more flexible point of view necessary for coding. If these perceptions are accurate, then it would seem that flexibility or conformity would be related to coding performance.

The training model is based on the assumption that the MTAI (as well as the MTAI-S) measures or is related to such personality variables as conformity. A study by Wakefield and Cunningham (1976) provides some empirical support for this assumption. These authors examined the relationship between five MTAI factors identified by Yee and Fruchter (1971) and Cattell's (Cattell, Eber, & Tatsuoka, 1970) Sixteen Personality Factors Questionnaire (16PF). Wakefield and Cunningham concluded that 41% of the variance of the five MTAI factors is shared by the 16PF. Thus, it is not unreasonable to speculate about personality variables such as conformity which might be measured by the MTAI (and the MTAI-S).

All of the above comments are, of course, quite speculative. The data collected for this study are not sufficient to discriminate between the attitude or training models or to eliminate other possible models which might be designed to understand coding performance. Future research is necessary to replicate the findings of this study. It would be useful if other predictors were included (perhaps measures of conformity or related personality variables) which could help pin down the crucial variables which account for variance in coding performance. Most helpful would be measures of subjects' reactions to the training process itself in order to understand in what way training interacts with coding performance. Perhaps one could measure subjects' resistance or conformity to the information and points of view expressed during coding. It would

be predicted that these measures would relate to final performance. Finally, finer-grained analysis of the coding process is necessary in order to fully understand it. It is difficult to surmise what processes underlie the coding task when it is analyzed in large time units of a half-hour or an hour. Code-by-code analysis would probably be most fruitful in understanding the quick decisions coders must make in time-based or interaction-based low-inference coding systems.

Reference Notes

- ¹The expert trainer and the training system were provided by EPIC Systems, 4901 East Fifth Street, Tucson, Arizona 85732.
- ²Haak, A., Lawrence, W.A., & Peck, R.F. ITEC: An experiment in school consultation. Publication No. 1082, The Research and Development Center for Teacher Education, The University of Texas at Austin, 1971.

References

- Cattell, R.B., Ebes, H.W., & Tatsuoka, M.M. Handbook for the Sixteen Personality Factor Questionnaire (16PF). Champaign, IL: Institute for Personality and Ability Testing, 1970.
- Cook, W.W., Leeds, C.H., & Callis, R. Minnesota Teacher Attitude Inventory: Manual. New York: Psychological Corp., 1951.
- Flanders, N.A. Analyzing Teaching Behavior. Reading, Massachusetts: Addison-Wesley, 1971.
- Marston, P.T., Zimmerer, L.K., & Vaughan, W.C. Coder drift: a reliability problem for teacher observations. Paper presented at American Educational Research Association, Toronto, Canada, 1978.
- Poynor, L.H., & Wetherold, R. Instrumentation for the evaluation of teaching project. Research Report No. 3, Research and Development Center for Teacher Education, The University of Texas at Austin, 1976.
- Rosenshine, B. Teaching behaviors and student achievement. London: International Association for the Evaluation of Educational Achievement, 1971.
- Scott, W.A. Reliability of content analysis: the case of nominal scale coding. The Public Opinion Quarterly, 1955, 19, 321-325.
- Strauss, G., & Sayles, L.R. Personnel: the Human Problems of Management. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1967.
- Wakefield, J.A., & Cunningham, C.H. Related factors of the MTAI and the 16PF. Psychology in the Schools, 1976, 13, 149-151.
- Winer, B.J. Statistical Principles in Experimental Design. New York, New York: McGraw-Hill, 1971.
- Yee, A.H., & Fruchter, B. Factor content of the Minnesota Teacher Attitude Inventory. American Educational Research Journal, 1971, 8, 119-133.

Table 1
Correlations Among FSIA Criterion
and Predictor Variables

<u>Variable</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>
1) Average reliab. across 11 tapes	----	.4460*	.4802**	.3488*	.4854**	-.3453*
2) MTAI-S		----	.3541*	.3305*	.4005*	-.0254
3) Training check 1			----	.4620**	.8576**	.0900
4) Training check 2				----	.8523**	.2388
5) Training checks 1 and 2					----	.1915
6) Experimental group						----

*p<.05, df=24

**p<.01, df=24

Table 2
Correlations of Four Indices of TES Average Reliability
Among Themselves and With the MTAI-S

<u>Correlations</u>	<u>PER</u>	<u>RXY</u>	<u>RIM</u>	<u>RIA</u>	<u>MTAI-S</u>
Percent agreement (PER)	---	.6315**	.7172**	.6413**	.3467*
Pearson Product-Moment correlation (RXY)		---	.8866**	.9978**	.1527
Unanchored intraclass (RIM)			---	.8957**	.1270
Anchored intraclass (RIA)				---	.1447
Means	75.75	51.57	45.55	49.35	
S.D.	3.93	6.54	7.22	6.40	

df = 24
*p<.05
**p<.01

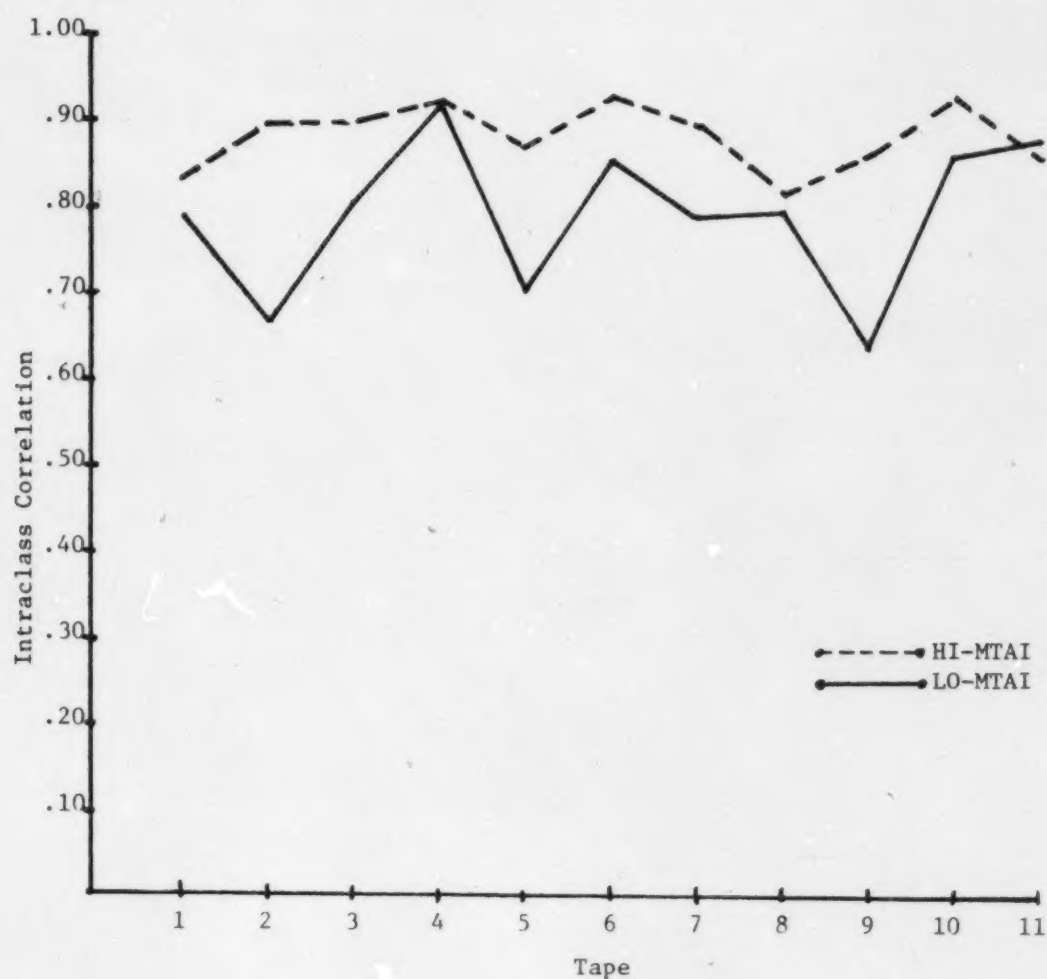


Figure 1. Intraclass correlations on each of the eleven experimental tapes for the HI-MTAI and LO-MTAI groups.

TEACHER EVALUATION SCALE
(TES)

1. Amount of Criticism (high = 1, low = 5) _____

Criticism, as defined for this instrument, refers to the behaviors a teacher uses in finding fault with and attempting to correct student cognitive processes and/or behavior.

2. Quality of Criticism (kind = 1, low = 5) _____

Kind teacher criticism is defined as being constructive, supportive, and considerate, achieving a positive effect. Harsh criticism has undertones of hostility, anger, and/or ridicule, producing a negative effect.

3. Amount of Praise and/or Encouragement (high = 1, low = 5) _____

Praise/Encouragement is defined as positive reinforcement of substantive or task related behavior of pupil.

4. Quality of Emotional Atmosphere (warm = 1, low = 5) _____

Warm atmosphere consists of teacher behavior which is kind, friendly, humanistic, and accepting of pupils' feelings. Cold atmosphere is demonstrated by aloofness and no acknowledgement of pupils' feelings.

5. Amount of Cognitive Variety (high = 1, low = 5) _____

Cognitive variety reflects the degree to which a teacher varies instructional modes to achieve a variety of cognitive processes in the class.

6. Amount of Response Elicitation (high = 1, low = 5) _____

Response-elicitation refers to the frequency with which teacher asks for verbal output from the pupil.

7. Amount of Lecturing (high = 1, low = 5) _____

Lecturing consists of giving factual information and explanations about content and procedure.

8. Amount of Direction Giving Behavior (high = 1, low = 5) _____

Direction giving behavior consists of commands and orders to which a student is expected to comply.

9. Amount of Opinion Giving (high = 1, low = 5) _____

Frequency with which teacher issues personal views or judgments.

Appendix B

This appendix contains the MTAI-S: the shortened version of the MTAI. In parentheses following each statement is its original item number on the MTAI (these numbers did not appear on the version of the questionnaire administered to subjects).

Classroom Attitude Questionnaire

This is a questionnaire about your attitudes toward classrooms, teachers, and pupils. Please indicate whether you agree or disagree with each of the statements -- there are no right or wrong answers. Use the following scale to mark the space next to each statement to indicate how you feel about it.

1 = Strongly Agree

3 = Undecided
or Uncertain

4 = Disagree

2 = Agree

5 = Strongly Disagree

- _____ 1. The teacher should not be expected to manage a child if the latter's parents are unable to do so. (15)
- _____ 2. Discipline in the modern school is not as strict as it should be. (17)
- _____ 3. Standards of work should vary with the pupil. (24)
- _____ 4. Increased freedom in the classroom creates confusion. (33)
- _____ 5. There is too much emphasis on grading. (35)
- _____ 6. There is too great an emphasis upon "keeping order" in the classroom. (37)
- _____ 7. There are times when a teacher cannot be blamed for losing patience with a pupil. (48)
- _____ 8. Too many children nowadays are allowed to have their own way. (53)
- _____ 9. Most pupils are resourceful when left on their own. (62)
- _____ 10. Children should be allowed more freedom in their execution of learning activities. (71)
- _____ 11. Pupils must learn to respect teachers if for no other reason than that they are teachers. (72)
- _____ 12. Difficult disciplinary problems are seldom the fault of the teacher. (77)
- _____ 13. The whims and impulsive desires of children are usually worthy of attention. (78)
- _____ 14. Children usually have a hard time following instructions. (79)
- _____ 15. Children are unable to reason adequately. (83)
- _____ 16. Teachers who are liked best probably have a better understanding of their pupils. (89)

(Adapted from the Minnesota Teacher Attitude Inventory and reproduced by permission for research purposes. Copyright 1951 by the Psychological Corporation, New York, New York. All rights reserved.)

1 = Strongly Agree
2 = Agree

3 = Undecided or
Uncertain

4 = Disagree
5 = Strongly Disagree

- ___ 17. Children should be given more freedom in the classroom than they usually get. (93)
- ___ 18. Pupils can be very boring at times. (98)
- ___ 19. Most pupils are considerate of their teachers. (101)
- ___ 20. A teacher should never leave the class to its own management. (105)
- ___ 21. "Lack of application" is probably one of the most frequent causes for failure. (108)
- ___ 22. Grading is of value because of the competition element. (112)
- ___ 23. Children usually will not think for themselves. (114)
- ___ 24. It isn't practical to base school work upon children's interests. (121)
- ___ 25. One should be able to get along with almost any child. (127)
- ___ 26. Children are not mature enough to make their own decisions. (128)
- ___ 27. Children will think for themselves if permitted. (130)
- ___ 28. Children should be given reasons for the restrictions placed upon them. (133)
- ___ 29. Most pupils are not interested in learning. (134)
- ___ 30. It is usually the uninteresting and difficult subjects that will do the pupil the most good. (135)
- ___ 31. Teachers should not expect pupils to like them. (141)
- ___ 32. Aggressive children require the most attention. (143)
- ___ 33. A pupil has the right to disagree openly with his teachers. (147)
- ___ 34. Most pupil misbehavior is done to annoy the teacher. (145)
- ___ 35. One should not expect pupils to enjoy school. (149)

Appendix C

The purpose of this appendix is to describe the data base and method used to examine the relationship between the MTAI-S and the MTAI.

Data Base

The data base consisted of 365 MTAI protocols collected from students seeking entry to a professional educational training program at the University of Lethbridge, Lethbridge, Alberta, Canada.¹ This program trains students in both elementary and secondary education. The average age of subjects in the sample was 23.6 (S.D.=4.6). Subjects normally took the test during their sophomore year, although some subjects were further advanced academically, a few having obtained Masters degrees. On the average, subjects had completed 14.3 (S.D.=6.4) semester courses at the university at the time at which they took the MTAI. The sample was approximately 51% female and 49% male.

Subjects were asked to take the MTAI as part of a battery of personality, aptitude, proficiency, and attitude tests. Subjects were told that these tests would be kept confidential and would be used for research purposes only. The goal of the research project was to develop models to predict student achievement in the teacher education program. Subjects were urged to complete the entire battery, but those with strong objections to the testing could be excused from it. Very few subjects were actually excused from the testing program, however.

Results

The 365 protocols obtained were examined for missing data. Those protocols containing missing responses were eliminated from consideration

¹Grateful acknowledgement for supplying the data used in this study is made to Dr. R. J. Twa, Associate Professor of Education, University of Lethbridge, who is a visiting scholar at the University of Texas Research and Development Center for Teacher Education for the 1977-78 academic year.

because the MTAI Scoring Manual (Cook et al., 1951) contained no guidelines for dealing with incomplete protocols. Data for 295 subjects remained after protocols with missing data were excluded.

These 295 protocols were scored in two ways. The first score obtained for each subject was the standard MTAI score based on all 150 items. The second score was based only on the 35 items included in the MTAI-S. In both cases, all items were scored in the manner prescribed by the MTAI scoring keys. The mean MTAI score was 41.32 (S.D.=31.78). This is slightly lower than the average score reported by Cook et al. for juniors in elementary and secondary education (mean=50.88, S.D.=28.02). The mean MTAI-S score was 9.44 (S.D.=8.23).

A Pearson product-moment correlation was computed between the MTAI and MTAI-S scores for the 295 subjects. A correlation of .8948 ($p < .01$) was obtained.

Discussion

The correlation between the MTAI and MTAI-S is sufficiently high that findings resulting from the use of the MTAI-S can most likely be replicated using the MTAI. The MTAI has been used in a large amount of research, and considering the strong relationship between the MTAI and the MTAI-S, the results of this MTAI research can be safely generalized to the MTAI-S. This fact may be useful in interpreting research using the MTAI-S.